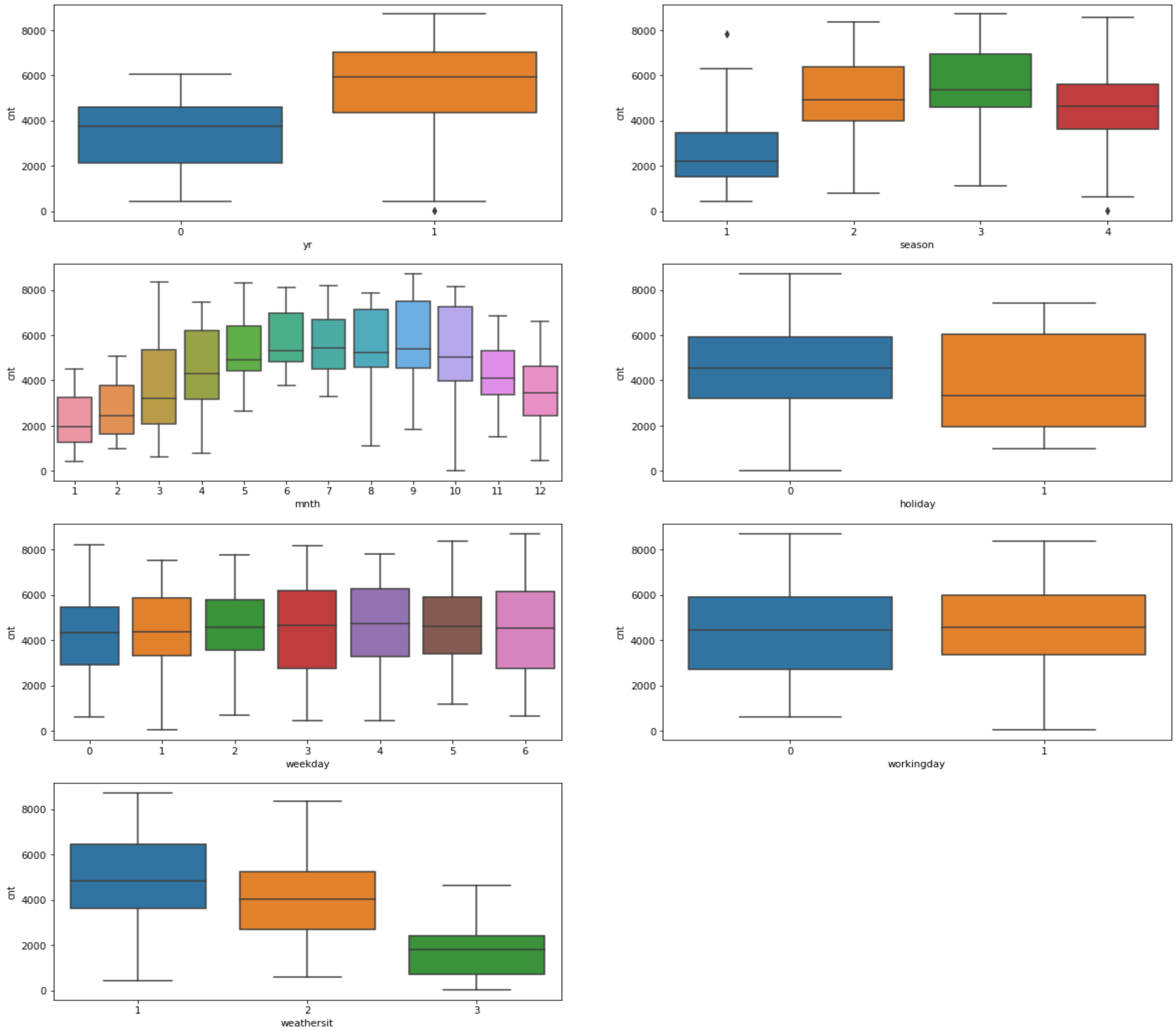
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)



Inferential analysis of categorical variables from the dataset on cnt is:

- **Yr**: Bike bookings are higher in 2019 as compared to 2018, it might be due to the fact bike rentals are getting popular and people are becoming more aware about environment.

- **season**: Highest booking happening in season3(fall) with a median of over 5000 booking. This was followed by season2(summer) & season4(winter) of total booking.

- **mnth**: Bike booking is quite high in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

- **weathersit**: Almost 67% of the bike booking were happening during ‘weathersit1 with a median of close to 5000 booking. This was followed by weathersit2. Clear weather is most optimal for bike renting.

- **holiday**: The bike booking was happening mostly when it is not a holiday.

- **weekday**: weekday variable shows very close trend. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

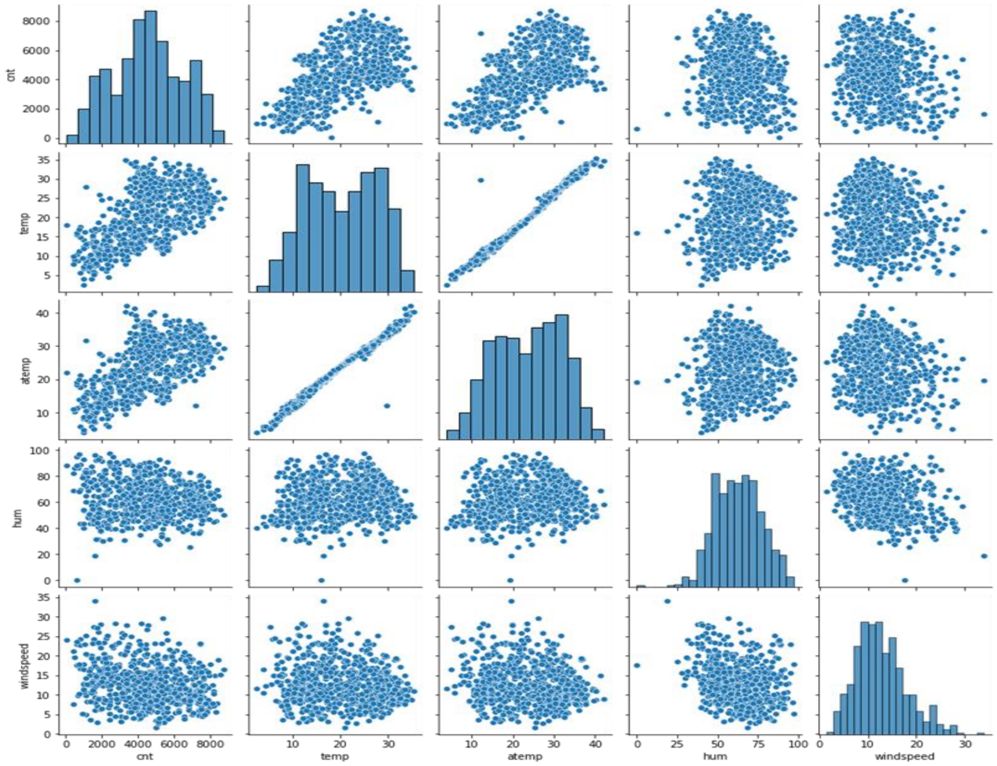
- **workingday**: Median is quite close, does not have much impact.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

If a field has three values, say "Red", "amber", "Green", when we do dummy, it creates three columns. Now statistically one of the column is redundant. Because we can arrive to the value using k-1 values itself.
Since the set of all k dummies creates multicollinear, we need to drop one.

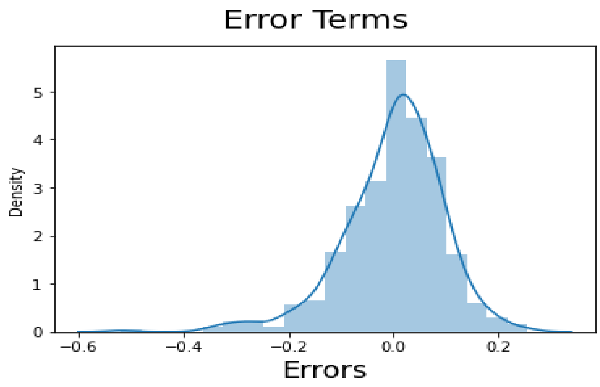
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

As we can see ‘temp’ and ‘atemp’ appear to highly corelated with target variable ‘cnt’

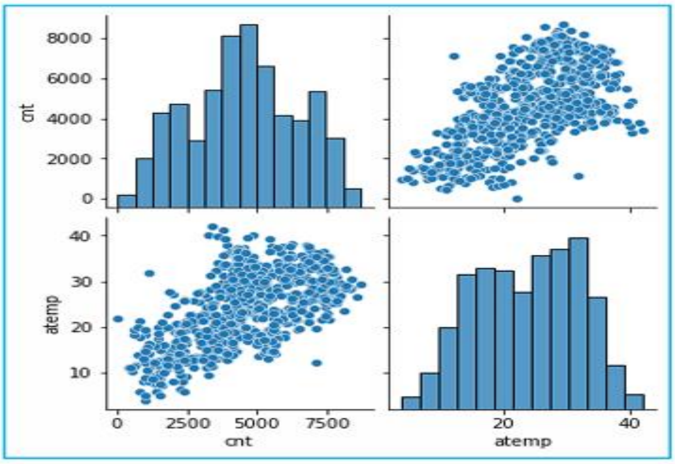


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

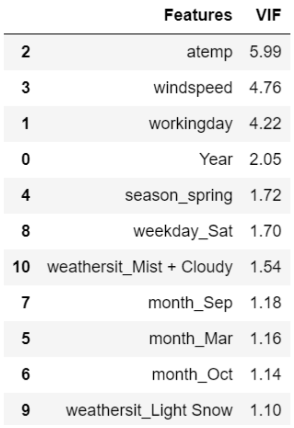
1. Error terms are normally distributed with mean zero.



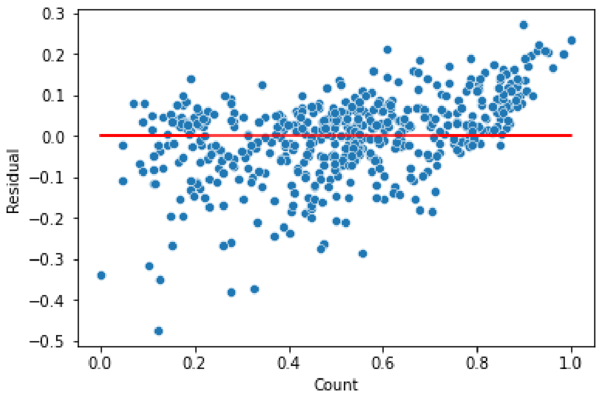
2. There is a linear relationship between X and Y



3. No Multicollinearity exists between the predictor variables. The values (other than the atemp) are well below 5.



	Features	VIF
2	atemp	5.99
3	windspeed	4.76
1	workingday	4.22
0	Year	2.05
4	season_spring	1.72
8	weekday_Sat	1.70
10	weathersit_Mist + Cloudy	1.54
7	month_Sep	1.18
5	month_Mar	1.16
6	month_Oct	1.14
9	weathersit_Light Snow	1.10



4. Residue error follow ‘Homoscedasticity’

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

- Temperature (atemp) - A coefficient value of '0.383568' indicated that a unit increase in temp variable increases the bike hire numbers by 0.383568 units.
- Year (yr) - A coefficient value of '0.240592' indicated that a unit increase in year variable increases the bike hire numbers by 0.240592 units.
- season_spring, windspeed, weathersit_light_snow, weathersit_Mist+Cloudy are negative coefficients which decrease the bike hire numbers.
- month_Mar, month_Oct and month_Sep - A coefficients value of 0.062384, 0.092330 & 0.084001 indicated that a (Mar, Oct & Sep) month, increases the bike hire numbers.
- Working_day, weekday_Sat - A coefficient value indicated that a working_day & weekday_Sat, increases the bike hire numbers.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

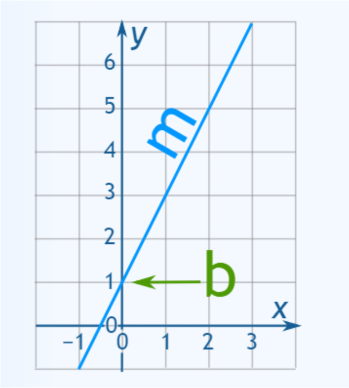
Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values.Linear Regression is the most basic form of regression analysis.Regression is the most commonly used predictive analysis model.

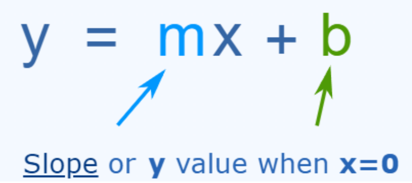
Linear regression is based on the popular equation **"y = mx + b".**

It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.

Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.





y = how far up

x = how far along

m = Slope or Gradient (how steep the line is)

b = value of **y** when **x=0**

Regression is broadly divided into simple linear regression and multiple linear regression.

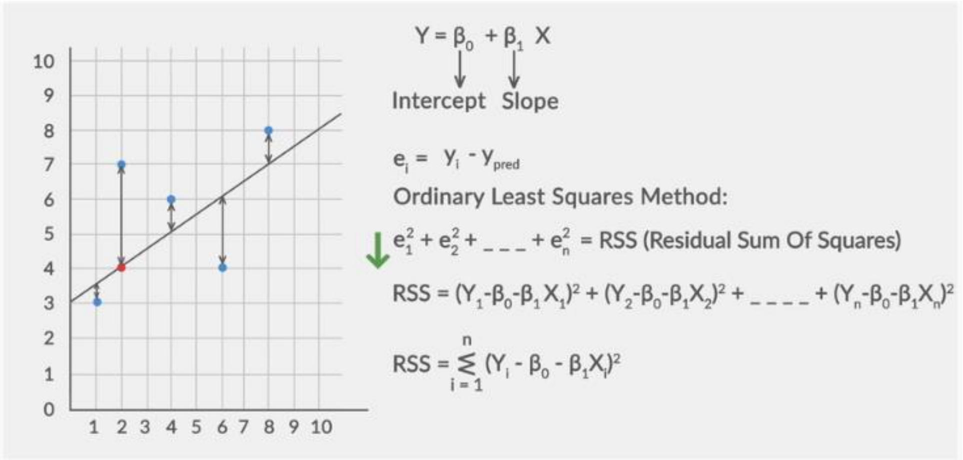
1. **Simple Linear Regression**: SLR is used when the dependent variable is predicted using only one independent variable.
2. **Multiple Linear Regression**: MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots ,$$

β1 = coefficient for X1 variable β2 = coefficient for X2 variable β3 = coefficient for X3 variable and so on… β0 is the intercept (constant term).

Once the line is fit, we have to find out whether the line is the best fit line using the RSS and TSS.



$$Y = \underset{\substack{\downarrow \\ \text{Intercept}}}{\beta_0} + \underset{\substack{\downarrow \\ \text{Slope}}}{\beta_1} X$$

$$e_i = y_i - y_{pred}$$

Ordinary Least Squares Method:

$$e_1^2 + e_2^2 + \_\,\_\,\_ + e_n^2 = \text{RSS (Residual Sum Of Squares)}$$

$$\text{RSS} = (Y_1 - \beta_0 - \beta_1 X_1)^2 + (Y_2 - \beta_0 - \beta_1 X_2)^2 + \_\,\_\,\_\,\_ + (Y_n - \beta_0 - \beta_1 X_n)^2$$

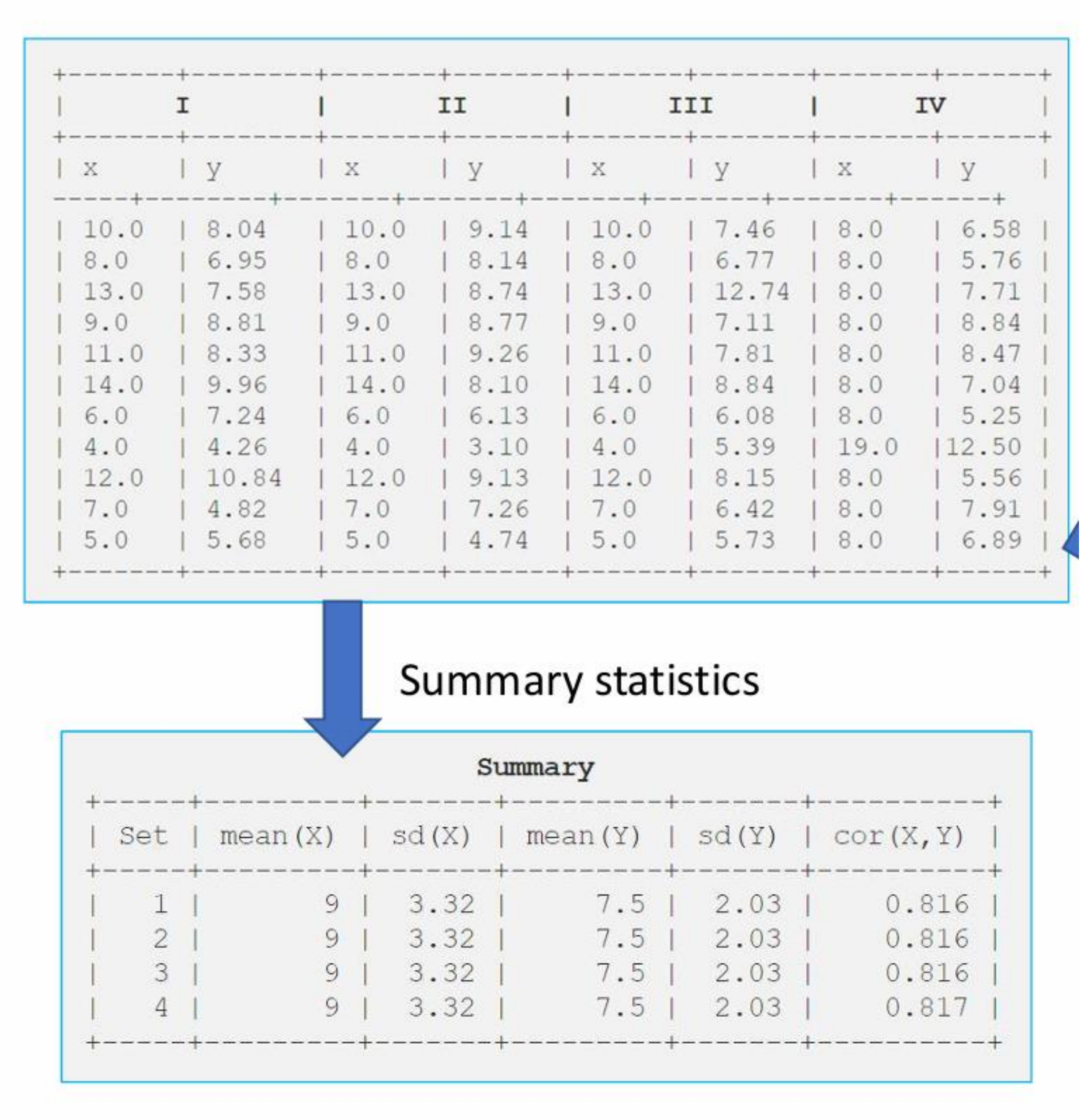
$$\text{RSS} = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

RSS : this is computed by considering the straight line, difference between the line and the actual data point, sq the diff and add them

TSS : this is computed by considering the avg of all data points, get diff, sq the diff, add them .

R^2 = 1 - (RSS/TSS)

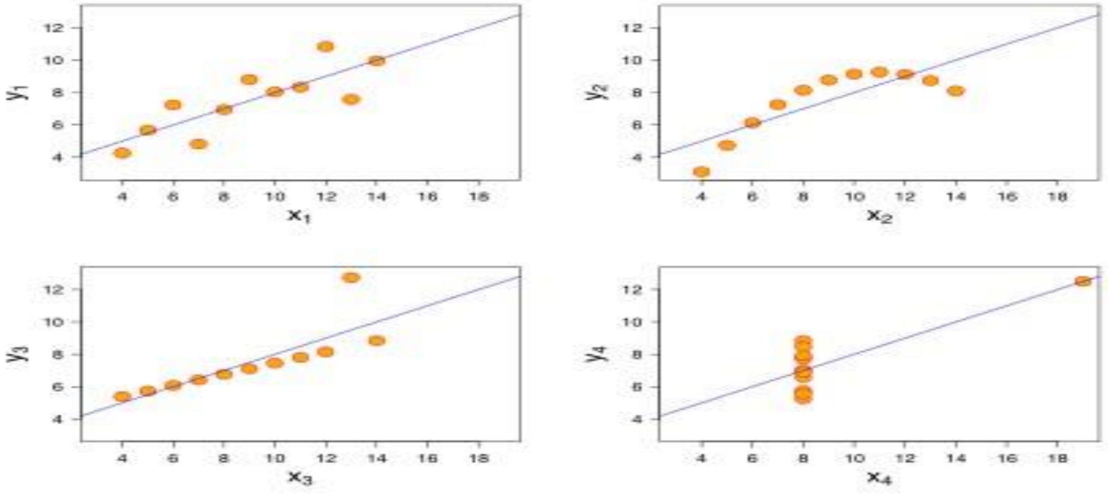
2. Explain the Anscombe's quartet in detail. (3 marks)



```
+-------+--------+-------+-------+-------+-------+-------+-------+
|       I        |      II       |     III       |      IV       |
+-------+--------+-------+-------+-------+-------+-------+-------+
| x     | y      | x     | y     | x     | y     | x     | y     |
-----+-------+-------+-------+-------+-------+-------+------+
| 10.0  | 8.04   | 10.0  | 9.14  | 10.0  | 7.46  | 8.0   | 6.58  |
| 8.0   | 6.95   | 8.0   | 8.14  | 8.0   | 6.77  | 8.0   | 5.76  |
| 13.0  | 7.58   | 13.0  | 8.74  | 13.0  | 12.74 | 8.0   | 7.71  |
| 9.0   | 8.81   | 9.0   | 8.77  | 9.0   | 7.11  | 8.0   | 8.84  |
| 11.0  | 8.33   | 11.0  | 9.26  | 11.0  | 7.81  | 8.0   | 8.47  |
| 14.0  | 9.96   | 14.0  | 8.10  | 14.0  | 8.84  | 8.0   | 7.04  |
| 6.0   | 7.24   | 6.0   | 6.13  | 6.0   | 6.08  | 8.0   | 5.25  |
| 4.0   | 4.26   | 4.0   | 3.10  | 4.0   | 5.39  | 19.0  |12.50  |
| 12.0  | 10.84  | 12.0  | 9.13  | 12.0  | 8.15  | 8.0   | 5.56  |
| 7.0   | 4.82   | 7.0   | 7.26  | 7.0   | 6.42  | 8.0   | 7.91  |
| 5.0   | 5.68   | 5.0   | 4.74  | 5.0   | 5.73  | 8.0   | 6.89  |
+-------+--------+-------+-------+-------+-------+-------+-------+
```

Summary statistics

```
                            Summary
+-----+----------+-------+----------+-------+-----------+
| Set | mean(X)  | sd(X) | mean(Y)  | sd(Y) | cor(X,Y)  |
+-----+----------+-------+----------+-------+-----------+
|   1 |        9 |  3.32 |      7.5 |  2.03 |     0.816 |
|   2 |        9 |  3.32 |      7.5 |  2.03 |     0.816 |
|   3 |        9 |  3.32 |      7.5 |  2.03 |     0.816 |
|   4 |        9 |  3.32 |      7.5 |  2.03 |     0.817 |
+-----+----------+-------+----------+-------+-----------+
```



My key takeaway from Anscombe’s quartet is that we have to visualize data using graphs, it is important to plot our data. Summary statistics alone is not sufficient.

Now looking at the data in the table, the summary statistics for all 4 sets it looks the same. But when we plot them as graphs, its all look totally different. So it proves how much Anscombe has visualized this data in his dream!

3. What is Pearson’s R? (3 marks)

Pearsons’s R measures the strength of the linear relationship between two variables.

Pearson’s R is always between -1 and +1

- The correlation coefficient lies between -1 and +1. *i.e.* -1 ≤ *r* ≤ 1
- A positive value of ‘*r*’ indicates positive correlation.
- A negative value of ‘*r*’ indicates negative correlation
- If *r* = +1, then the correlation is perfect positive • If *r* = –1, then the correlation is perfect negative.

If *r* = 0, then the variables are uncorrelated.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Feature **scaling** is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

- **Standardized_val = (input_value – mean) / standard_deviation**
- **MinMax_val = (x–xmin)/ (xmax–xmin)**

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

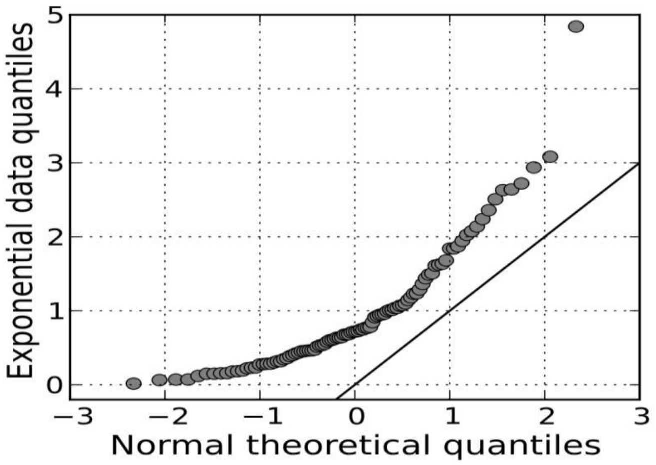
VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. (VIF) =1/(1-R^2). If there is perfect correlation, then VIF = infinity. Where R^2 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1.So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity" **.**

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?


Exponential data quantiles
Normal theoretical quantiles